A Bayesian Critique of the Likelihood Principle
Discussion of The Likelihood Principle: A Review, Generalizations, and Statistical Implications by James Berger and Robert Wolpert

by

David A. Lane

Department of Theoretical Statistics
University of Minnesota
Technical Report No. 432

Introduction

Berger and Wolpert have done the statistics community a service by calling our attention once again to the likelihood principle (LP) and its implications. They repeat Birnbaum's (1962) message, already admirably recapitulated by Basu (1975) and Dawid (1977): if you work within the classical $(X, \theta, \{P_\theta\})$-paradigm, you want to make inferences about "true θ" on the basis of "observed x," and you wish to respect certain fundamental principles of inference (for example, the sufficiency and weak conditionality principles), then your inference had better depend upon the observation x through the likelihood function that x induces on θ. In particular, you must accept the implications of some other principles that many statisticians regard as false, never mind fundamental, like the stopping time and censoring principles.

There are several bail-out options for statisticians who choose neither to follow the LP to fully conditional analysis nor to raise adhockery to a scientific principle. They can reject the $(X, \theta, \{P_\theta\})$-paradigm by requiring either more structure (as do structuralists, pivoteers, and, perhaps, some "objective" Bayesians) or less (as do defenders of alternative-free significance tests and, more drastically, exploratory data analysts); or they can modify the fundamental pre-principles so that the LP and the objectionable post-principles fail to be derivable from them, as did Durbin (1970) and Kalbfleisch (1975); or they can claim that other, more fundamental principles, like

the Confidence Principle, conflict with the LP, making an ideological choice among competing principles necessary.

Since Bayesian practice is consistent with the LP, Bayesians have no need to refute Birnbaum's work. Indeed, to Berger and Wolpert, the LP is a trump card in the Bayesian salesman's hand. They argue, as did Basu (1975), that only Bayesian ideas permit the LP to be properly implemented and that Bayesian considerations unravel the "counterexamples" to the LP produced by Armitage, Stein, Fraser and others.

But even to Bayesians, consistency with Bayesian ideas should be no guarantee of foundational cogency. For example, the fact that (essentially) admissible decision rules are Bayes does not recommend Wald's formulation of decision theory to most Bayesians. So the question arises: should Bayesians promote Birnbaum's formulation and derivation of the LP as a cornerstone of the foundations of statistics? I think not, for two reasons. First, the LP is embedded in a paradigm which is not directly applicable to many, if not most, of the important real problems of statistical inference. Because of the ambiguity and limitations of this paradigm, the proof of the LP is not compelling. Second, the LP ignores what I regard as the fundamental tenet of Bayesianity: the purpose of an inference is to quantitate uncertainty. When this tenet is properly taken into account, foundational arguments can be adduced that lead directly to Bayesian methods.

The next section elaborates the first of these reasons in some detail. For a development of the second, see Lane (1981)

and Lane and Sudderth (1984).

$(X, \Theta, \{P_\theta\})$ and LP

I shall discuss three problems with the LP. The first relates to the meaning, the second to the adequacy, and the third to the relevance of the $(X, \Theta, \{P_\theta\})$-paradigm. Both the first and the second of these problems call the derivation of the LP into question.

1). What do the elements of Θ represent? This question is important, since the proof of the LP requires us to consider the mixture of two different experiments with "the same Θ." There are at least three possible interpretations of the elements Θ:

a) Θ <u>is</u> the distribution P_θ;

b) Θ is an abstract set and Θ merely indexes the distribution P_θ;

c) Θ is a possible value for some "real" physical parameter, and P_θ is to be regarded as the distribution of the random quantity x should Θ be the true value of that parameter.

Interpretations a) and b) are mathematically precise. They are defined in terms of the assumed model and do not refer to the physical reality that model is intended to represent.

Interpretation c) has an entirely different character and

raises difficult philosophical issues. When - and in what sense - do "real" physical parameters exist? If I opt for interpretation c), must I believe that a coin has a propensity to come up heads $\theta \times 100\%$ of the time in an (infinitely) long series of repeated flips? I am inclined to believe that there may be "real" physical parameters in measurement error problems, although even here a strict operationalist construction leads to interpretation a) rather that c) for the parameter θ: the measuring process, encoded as P_θ, defines the quantity measured. In few other problems to which statistical inference is applied are there model-free physical quantities standing behind each model parameter. To decide whether or not you agree, think about your last regression or time-series analysis.

Both Berger and Wolpert (pp. 42-3) and Dawid (1977, P. 252) seem to favor interpretation c). For example, Berger and Wolpert say that the LP applies only when the elements of the two parameter sets are "the same parameter, i.e. are physically or conceptually the same quantity." Unfortunately, they neglect to tell us how we are to decide when two different experiments measure the same quantity or how to deal with model parameters that lack any natural interpretation in terms of physical quantities. Moreover, in virtually all of their examples the set θ is uninterpreted and merely serves to index the set of distributions $\{P_\theta\}$, which suggests that in these cases they are thinking about θ in the sense of interpretation b). It is hard to take the LP seriously as a foundational instrument if we must always interpret the elements of θ as "real" physical quantities,

unless we are given some guidance on what constitutes reality and how reality is tied to mathematics by the model we select.

It matters which of the three interpretations we give to the elements of θ. They lead to very different conclusions about the validity of the derivation of the LP. Interpretation a) gives no scope for the mixture principle: only experiments whose sampling distributions are identical share "the same θ." As such, the LP is reduced to the sufficiency principle and; for example, the stopping time principle does not follow from the LP.

Interpretation b), on the other hand, gives tremendous scope for mixing. Any two experiments with the same index set can be mixed. Consequently, if there are a pair of observations, one from each experiment, that yield the same likelihood function on the index set θ, the LP then declares that the "evidence" or "inference" derived from the two experiments with these two observations must be identical. This is a startlingly unBayesian conclusion. For example, must my predictive inference for the next outcome in <u>any</u> sequence of Bernoulli trials in which I have so far obtained three successes and one failure be the same? But what in the mathematics of the LP proof precludes interpreting θ purely as an index set and so deriving a version of the LP that conflicts with Bayesian practice?

The foundational status of the LP cannot be determined until θ is interpreted. Depending on whether one adopts interpretation a), b) or c), the LP is devoid of interesting consequences, wrong, or severely and ambiguously restricted in its domain of applicability.

2). The proof of the LP is convincing only in so far as the sufficiency and weak conditionality principles are intuitively compelling. While Bayesian practice respects both principles, only weak conditionality seems unarguable on its face. I share I.J. Good's reaction to the sufficiency principle, as reported in his discussion of Birnbaum (1962). Despite Fisher's gift for suggestive names (what more could you possibly need than something that is sufficient?), the fact that the distribution of x given the value of a statistic T is θ-free does not immediately impel me to base my inference only on the value of T.

Suppose, though, that the observation x is generated by first generating a value for T according to a distribution indexed by some element of the parameter set θ, and then an extraneous randomization mechanism is used to pick an x on the orbit of the observed value of T. In such a case, it is clear that T is sufficient and that inference about θ should be based only on T. (The sufficient statistic that appears in the derivation of the LP does not bear this postrandomization relation to the observation x.)

Now for any sufficient statistic T defined on a statistical model $(X, \theta, \{P_\theta\})$, there is no way to tell from the information encoded in $(X, \theta, \{P_\theta\})$ whether the observation x is or is not generated from T by postrandomization. So, if you do not find the sufficiency principle compelling except in the postrandomization case, you must agree with Barnard and Fraser that not enough information is encoded in $(X, \theta, \{P_\theta\})$ upon which

to base a general principle of inference. And I believe that this conclusion is correct. After all, the information in $(X, \theta, \{P_\theta\})$ says nothing about how the model represents reality, and it is hard to see how a principle of inference can disregard the details of this representation. Though we use models to guide the way we formulate inferences, the inferences themselves have value to us only if they yield useful statements about the world.

3). Even though "inference" is undefined in the LP formulation, the validity of the LP seems to depend on two premises about the nature of inference in the $(X, \theta, \{P_\theta\})$-paradigm:

a) the purpose of inference is to make some statement about the "true" value of an unobservable parameter θ on the basis of an observed quantity x;

b) θ exists independently of the "experiment" E that produces x, and information about θ can be separated into two components, one deriving just from E (to which the LP refers) and the other from "other information" presumably preexisting E.

I believe that these premises are rarely true in real situations to which statistical inference is applied. If I am right, the scope of the LP as a foundational instrument is narrow.

Except for measurement error problems, the real aim of inference is usually to generate a prediction about the value of some future observables; see Geisser (1971 and 1984) and

Aitchison and Dunsmore (1975) for extensive discussion of this proposition and further references. This is especially true in situations where the model parameters do not represent real physical quantities, the typical case in regression and time-series analyses. Estimating model parameters is in general a "half-way house" on the way to predicting some relevant future observation, and much can be lost by focussing foundational discussion on the half-way house instead of the ultimate destination. For example, the relevant uncertainty for a patient with a particular clinical condition undergoing a particular therapy is not a confidence band for an estimated survival curve; rather, the patient and his physician should be concerned with the predictive distribution for that patient's future lifetime. The inferential question of interest to the patient is how to generate this predictive distribution.

The LP does not address this question directly. Berger and Wolpert claim that prediction can be embedded in the LP framework by including the future observable as part of the unknown parameter. But then θ itself appears as a nuisance parameter that is clearly not "noninformative" in the sense of Berger and Wolpert. LP ideas provide no guidance on the treatment of informative nuisance parameters. On the other hand, deFinetti's subjective Bayesian theory is directed towards the problem of predicting future observables, and the notion of coherence derived from that theory provides a foundational basis for predictive inference; see Lane and Sudderth (1984). In this theory, models may be used to help generate predictions about the

future observable y based upon observed x, but the models merely provide a convenient structure and need carry no metaphysical burden of "reality" for the parameters they contain.

The premise b) cited above ignores the fact that model parameters are frequently inseparable from the "experiment" whose possible distributions they index. Especially in applications arising in nonexperimental sciences like econometrics or resource management, the model is sculptured either from data already in hand or perhaps from a realistic view of what data are potentially obtainable. In such cases, there is no way to separate what (E,x) says about θ from "prior" information about θ; in fact, θ cannot be said to exist prior to the information of E, even though there may be much prior information about which x might be observed. In these situations it is hard to criticize "objective" Bayesians who violate the LP by letting their "priors" depend upon the structure of the experiment E.

References

Aitchison, J. and I.R. Dunsmore (1975), Statistical Prediction Analysis (Cambridge Univ. Press).

Basu, D. (1975), "Statistical Information and Likelihood," Sankaya Ser. A 37, 1-71.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference," JASA 57, 269-306.

Dawid, A.P. (1977), "Conformity of Inference Patterns," in Recent Developments in Statistics, J.R. Barra et. al eds, (North-Holland).

Durbin, J. (1970), "On Birnbaum's Theorem on the Relation between Sufficiency, Conditionality, and the Likelihood," JASA 65, 395-398.

Geisser, S. (1971), "The Inferential Use of Predictive Distributions," in Foundations of Statistical Inference, V.P. Godambe and D.A. Sprott, eds. (Holt, Rinehart and Winston).

Geisser, S. (1984), "On the Prediction of Observables: A Selective Update," in Bayesian Statistics II, J.M. Bernardo et. al., eds.

Kalbfleisch, J.D. (1975), "Sufficiency and Conditionality," Biometrika 62, 251-268.

Lane, D. (1981), "Coherence and Prediction," Proceedings of the 43rd Session ISI, 81-96.

Lane, D. and W. Sudderth (1984), "Coherent Predictive Inference," Sankhya, to appear.